DOCUMENT RESUME

ED 346 120 TM 018 368

AUTHOR Schultz, Matthew T.; Geisinger, Kurt F.

TITLE The Effects of Sample Size and Matching Strategy on Mantel-Haenszel and Logit DIF Procedures.

PUB DATE Apr 92

NOTE 26p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (San Francisco, CA, April 1992).

PUB TYPE Reports - Descriptive (141) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.

DESCRIPTORS College Entrance Examinations; Comparative Analysis; Control Groups; Equations (Mathematics); Evaluation Methods; Experimental Groups; High Schools; High School Students; *Item Bias; Item Response Theory; *Mathematical Models; Minority Groups; *Research Methodology; Research Problems; *Sample Size; Sampling; *Test Items

IDENTIFIERS *Logit Analysis; *Mantel Haenszel Procedure; Matching to Sample Procedure; Scholastic Aptitude Test

ABSTRACT

Research efforts have established that the Mantel-Haenszel procedure (MHP) is an effective method for detecting the presence of test items exhibiting differential item functioning (DIF). While the MHP has been advocated for situations where item response theory based methods may not be usable, recent findings have suggested that the performance of the MHP and Logit needs to be examined in detail. The present research examined the impact of manipulating sample size, the ratio of focal to reference group members, and the number of levels of the matching criterion on the performance of the MHP and Logit. Reference and focal groups consisted of 7,320 majority group Scholastic Aptitude Examination (SAT) takers and 791 minority group SAT takers. Other samples of 1,000 reference and 500 focal group members were obtained through a random number generating program. Four, 8, and 12 levels of matching on the SAT score were used. Results suggest that the MHP and Logit are sensitive to these manipulations. As sample sizes decreased, mean values for the MHP and Logit decreased, and agreement between them declined. As the number of levels of matching increased in the full sample condition, agreement between Logit and the MHP also dropped. Four tables present study findings, and there is a 23-item list of references. (Author/SLD)

The Effects of Sample Size and Matching Strategy on Mantel-Haenszel and Logit DIF Procedures

Matthew T. Schultz

Law School Admission Services

and

Kurt F. Geisinger

Fordham University

Paper presented at the Annual Meeting

of the National Council on Measurement in Education

San Francisco, CA, April 1992

Abstract

Research efforts have established that the Mantel-Haenszel (MH) procedure is an effective procedure for detecting the presence of test items exhibiting differential item functioning (DIF). While MH has been advocated for situations where IRT-based methods may not be usable, recent findings have suggested that the performance of MH and Logit needs to be examined in detail. The present research examined the impact of manipulating sample size, the ratio of focal to reference group members, and the number of levels of the matching criterion on the performance of MH and Logit. Results suggest that MH and Logit are sensitive to these manipulations.

In the past twenty-five years considerable research efforts have examined the existence of possible ethnic, racial or gender group differences in the predictive use of test scores. The increasing use of tests for educational and industrial evaluation, assessment, selection and placement has resulted in heightened awareness of group differences in performance and a demand for demonstrated test validity and fairness.

A test item is considered biased, or to exhibit differential item functioning (DIF) it individuals with similar ability, but from different groups, have a different probability of answering an item correctly. A typical finding is that, on any given exam, a number of items will be found to be differentially difficult for members of disadvantaged groups.

While the above definition of DIF is generally accepted, there is no such consensus regarding the selection of a particular procedure for identifying such items from among the large number of procedures that have been advocated. One procedure that has generally been shown to perform well is the Mantel-Haenszel (MH) procedure, which was developed originally within the context of biomedical research (Mantel and Haenszel, 1959). More recently, MH has been adapted for the detection of DIF by Holland and Thayer (1988). The Logit procedure (Mellenbergh, 1982) is computed similarly to MH and should offer similar results. For the analysis of DIF, both procedures match for ability on some criterion, generally total test score (or total test score reflecting

adjustments due to the removal of items demonstrating DIF). Differences beyond this adjustment are interpreted as evidence of DIF. These procedures are attractive alternatives to more computationally cumbersome procedures (e.g., IRT-based techniques) because they are easy both to compute and interpret.

The Logit Procedure

The following description of models for the three-dimensional table used in DIF research is based upon Mellenbergh (1982) and reflects categories for test score intervals, group and item responses. The expected frequency in the i score category ($i = 1, 2, \ldots, s$), j group ($j = 1, 2, \ldots, g$), and k response category where $k = 1$ for a correct response and $k = 0$ for an incorrect response, is denoted F_{ijk}, respectively.

The logit is defined as the natural logarithm of the ratio of correct and incorrect responses. The saturated logit model is (Fienberg, 1980):

$$\ln (F_{ij1}/F_{ij2}) = C + S_i + G_j + (SG)_{ij} , \qquad (1)$$

where ln denotes the natural logarithm and the constraints are as follows:

$$\sum_{i=1}^{s} S_i = 0, \qquad (2)$$

$$\sum_{j=1}^{G} G_j = 0, \qquad (3)$$

$$\sum_{I=1}^{S} (SG)_{ij} = \sum_{I=1}^{G} (SG)_{ij} = 0 \quad . \quad (4)$$

The parameters in logit can be interpreted similarly to parameters in an ANOVA model: C is the overall item difficulty parameter, S_i is the main effect for total score category, G_j is the main group effect, and $(SG)_{ij}$ is the score category by group interaction effect parameter. F_{ijk} is interpreted as defined above. For DIF analysis, two other, non-saturated logit models are also of interest (Mellenbergh, 1982). Deleting the interaction parameter yields:

$$\ln(F_{ij1}/F_{ij2}) = C + S_i + G_j, \qquad (5)$$

with the constraints given in 2 (the effect for score = 0) and 3 (the effect for group = 0). Deleting the group paramet yields:

$$\ln(F_{ij1}/F_{ij2}) = C + S_i \ , \qquad (6)$$

with the constraint given in 2.

If either model (equation) 1 or 5 is needed to describe the data, the item is biased. Model fit is assessed by computing the expected frequencies given the model and the likelihood ratio, which is asymptotically chi-square distributed. The expected frequency for model 5 is computed as follows:

$$\hat{f}_{ijk} = (\sum_{j=1}^{J} f_{ijk})(\sum_{k=1}^{2} f_{ijk}) / (\sum_{j=1}^{J}\sum_{k=1}^{2} f_{ijk}), \qquad (7)$$

where f_{ijk} is the sample observed frequency. The likelihood ratio statistic is

$$G^2 = 2\sum\sum\sum f_{ijk} \ln(f_{ijk}/\hat{f}_{ijk}), \qquad (8)$$

which is asymptotically chi-square distributed with s(g-1) degrees of freedom.

The Mantel-Haenszel Procedure

The first step in calculating MH is the selection of a criterion of ability on which to match examinees. Generally total test score is used (Holland & Thayer, 1989), however external criteria may be substituted. After selection of a matching criterion, the data for reference (majority) and focal (minority) group members are formed into a 2 by 2 by k table, (where there are two levels of performance on the item, right and wrong; two levels of grouping, reference and focal; k levels of matching on the criterion, where k refers to the number of matching groups) for each item j. Table 1 presents a sample MH format.

Insert Table 1 about here

In Table 1, A_j represents reference group members answering the item correctly and B_j reference group members answering incorrectly. C_j and D_j convey parallel information for the focal group. nR_j and nF_j denote the number of reference and focal group members, respectively, in the i^{th} matched group (by total score, for example), while T_j denotes the total number of examinees in the matched group. R_j and W_j denote the number of examinees answering the item correctly and incorrectly.

The test statistic of MH is given by the formula:

$$\chi^2_{MH} = \frac{(|\sum_1^J A_j \sum_1^J E(A_j)| - .5)^2}{\sum_1^J Var(A_j)} \qquad (9)$$

where $E(A_j) = n_{Rj}R_j/T_j$, and

$$\sum_1^J Var(A_j) = \frac{n_{Rj}n_{Fj}R_jW_j}{T_j^2(T_j - 1)} . \qquad (10)$$

An alternate statistic, $\hat{\alpha}_{MH}$ (alpha) is also frequently calculated, and gives the average factor by which the odds that a reference group member correct exceeds the corresponding odds for comparable focal group members. $\hat{\alpha}_{MH}$ is calculated using the following formula:

$$\hat{\alpha}_{MH} = \frac{\Sigma A_j D_j / T_j}{\Sigma B_j C_j / T_j} \qquad (11)$$

The focal group has the advantage when $\hat{\alpha}_{MH}$ is greater than one, whereas the converse is true when it is less than one. Values of zero correspond to items where no group is advantaged.

Comparative Research

MH, and to a lesser extent, Logit, have both been contrasted favorably to other procedures for detecting DIF. The most common comparison has been of MH with an IRT-based procedure. While the MH test statistic is not as widely utilized as the parameter estimate α_{MH} as a measure of DIF, other researchers (Donoghue & Allen, 1991; Shermis & St. George, 1990) have found that the MH X^2 statistic is a satisfactory measure of DIF. Hambleton and Rogers,

(1988), Hambleton, Rogers and Arrasmith (1988), Schulz, Perlman, Rice and Wright (1989), Perlman, Bezruczko, Reynolds, Rice and Schulz (1988), and Raju, Bode and Larsen (1989) each contrasted MH to a variant of the IRT procedures and found that MH frequently performed nearly as well at substantially lower cost. Camilli and Smith (1990) compared MH to the randomized and jackknifed test and found that MH performed quite well. Van Der Flier, Mellenbergh, Ader and Wijn (1984) and Kok, Mellenbergh and Van Der Flier (1985) similarly found that the Logit procedure is quite effective as identifying items exhibiting DIF.

As noted above, the MH procedure, and to a lesser extent the Logit procedure, have been compared favorably to Rasch and IRT-based procedures in a number of comparative studies. A disadvantage to these procedures is that the procedures susceptibility to changes in sample size and number of matching levels has only of late been studied in detail, as is described below.

Research Examining Manipulations of Sample Size and Matching Levels

Research to date has suggested that MH may be somewhat sensitive to manipulation of both sample size as well as changes in the number of levels of the matching criterion. In general, as sample sizes increase, the number of flagged items increases well. Mazor, Clauser and Hambleton (1991) examined the ability of MH to detect DIF under conditions of 100, 200, 500, 1000 and 2000 examinees per group for tests consisting of 75 items. As the number of examinees per group decreased, the ability of MH to

detect DIF dropped from a high of 69% of the items correctly identified to a low of 13%

Research examining the number of levels for matching has yielded some general guidelines. Raju, Bode and Larsen (1989) suggested a minimum of 4 levels, while Wright (1986) found that 61 levels were better than 6. The larger issue, as noted by Bradley, Fitzpatrick, and Sykes (1991) is determining what is the best operational criteria for adequate measurement of the matching criteria. This problem is especially critical considering MH's assumption that ability is held constant. Bradley et al. (1991) examined 13 levels (50 examinees per group), 14 levels (25 examinees), 9 levels (20 examinees), 12 levels (15 examinees), and 19 levels (10 examinees) for a test consisting of 299 items. There was a general tendency for MH alpha to remain stable as the minimum number of examinees dropped from 50 to 10. Donoghue and Allen (1991) examined tests consisting of 5, 10, 20 and 40 items using sample sizes of 400, 800, and 1600 examinees, 75% of whom were majority (reference) group members and 25% minority (focal) group members. Matching strategies included "thin" (one level for each possible raw score) as well as several "thick" matching approaches. For short tests (5 to 10 items) thin matching worked poorly. For longer tests (40 items) with large sample sizes (1600), both thin and several variants of thick matching worked well, though thin matching worked best. The MH measure used was found to influence results, with χ^2 performing best in conditions where equal number of group members were pooled. Equal interval methods were not

assessed in detail.

The purpose of the present research was to continue assessing the influence of 1) differing matching approaches and 2) manipulating the ratio of reference to focal group members when MH and Logit are applied to test data.

Method

The present study employed a national sample of 10,410 examinees who took the November, 1989 form of the SAT. The reference and focal groups consisted of 7320 majority group test takers and 791 minority group members as identified on a self-reported questionnaire, respectively. No attempt was made to control for gender (While two separate minority groups were utilized as focal groups, only results based upon the larger group are included in the present paper). The SAT consists of two sections, verbal (SAT-V) and mathematical (SAT-M), containing a total of 145 multiple choice items. Possible scores on each section range from 200 to 800.

MH χ^2 and Logit G^2 were calculated for each item across the following conditions:

	Reference	Focal
Full Sample	7320	791
Sample 1	1000	500
Sample 2	500	500

The full sample group consisted of all members from the national sample belonging to the groups designated reference and

focal. The reduced samples were obtained by utilizing a random number generating program (SPSS-X, 1988) to drop cases. Hence, samples 1 and 2 are samples taken at random from the full sample condition. It should be noted that for the focal group only, the sample 1 and 2 conditions contain the same sample of test takers. Thus, the only difference between these conditions is that the majority group has been sampled down to a size equivalent to that of the focal group.

Matching was done using the total subtest score as the matching criterion. Three samples were generated, using 4, 8 and 12 levels of matching on SAT score. It should be noted that because of range restriction (the focal group had few test takers with total scores at the upper end of the scale), the highest points in the 8 and 12 level conditions are quite broad. For four levels of matching, the score ranges 200-290, 300-390, 400-490, and 500-590 were utilized; For eight levels, the ranges were 200-250, 260-290, 300-340, 350-390, 400-440, 450-490, 500-540, and 550-700; for twelve levels, the ranges were 200-250, 260-280, 290-310, 320-340, 350-370, 380-400, 410-430, 440-460, 470-490, 500-520, 530-550, and 560-700. These bands were determined so that the number in each band were roughly similar.

Results

Kappa coefficients (Cohen, 1960; Siegel and Castellan, 1988) were computed to assess the degree of agreement across procedures, samples and levels. Table 2 presents the Kappa values found in classifying items as biased or not as a function of sample size.

Insert Table 2 about here

Overall, MH procedures demonstrated agreement with each other regardless of the number of levels of matching utilized (4, 8 or 12). The relationship between MH and Logit was not as clear; in the full sample conditions MH and Logit demonstrated significant agreement. However, as the sample sizes (and disparity between groups in terms of sample size) was reduced, agreement between Logit and MH decreased. Table 3 presents descriptive statistics for MH and Logit. MH χ^2 decreased both as sample size did and as the number of matching levels increased.

Insert Table 3 about here

Table 4 contains the Pearson product-moment correlations between procedures across samples.

Insert Table 4 about here

Discussion

Overall, the MH and Logit procedures demonstrated consistently high agreement with one another in the full sample condition, suggesting that the procedures may be used interchangeably with

large samples. The ratio of reference to focal group members (approximately 10 to 1) was considerably higher than Donoghue and Allen's (1991) 3 to 1 ratio, however results were parallel, in that the value of χ^2 increased as sample sizes did. As sample size decreased (and the ratio of focal to reference group members approached 1 to 1), agreement between MH procedures tended to increase, while agreement between MH procedures and Logit tended to decline substantially. In addition, the mean levels, and consequently the number of items identified as exhibiting DIF, tended to decrease as sample sizes did, supporting findings by Mazor et al. (1991) and Wise (1987) suggesting that MH tends to miss differentially functioning items as the number of examinees decreases. While this finding alone is not surprising due to the concomitant loss of power as samples become smaller, it is worthy of note because MH and Logit have been offered as procedures relatively insensitive to sample size considerations, and hence especially appropriate for conditions with small numbers of examinees. The general finding was that while fewer items were labeled "biased" as sample sizes decreased, the relationship among indices and items remained constant.

As the number of levels of matching increased from four to twelve, mean MH χ^2 values decreased. This finding supports suggestions by Raju et al., (1989), Schulz et al. (1989), and Wright (1986) that more levels of matching are better than fewer. In addition, the finding that MH values decrease as number of matching levels increases (and hence becomes finer) is consistent

with some current testing practices, which result in some testing organizations utilizing as many levels of matching as there are items. It has also been noted that as the number of levels of stratification on the matching criterion increases, MH findings will parallel Rasch (Schulz et al., 1989). The finding that reducing the ratio of reference to focal group members from 10 to 1 to 1 to 1 may impact on findings is a potential concern. The reduction in both mean levels of the MH χ^2 and Logit G^2 and number of items identified may be attributable to decreasing power or to instability in the procedures themselves.

The finding that correlations between procedures were relatively higher than agreement in classification as biased or not suggests that researchers must remain aware of the impact of reliance on a statistical significance test for determining the quality of an item. Correlations between MH-4, 8 and 12 were somewhat lower than those obtained by Raju et al., (1989) (lowest .998), which may be due to the multidimensional nature of the SAT. Correlations between Logit and MH procedures (which have not been empirically contrasted before) revealed substantial agreement, suggesting that these procedures do indeed function similarly. The finding of agreement between these procedures is consistent with Holland's (1985) observation that the two should provide near-identical results. In terms of rank order of magnitude of indices, the procedures clearly are more similar than when assessing their agreement in classifying items (via Kappa coefficients) as exhibiting DIF.

The lack of agreement between MH and Logit is also worth noting. While in the full sample condition Logit with 4 levels demonstrated considerable agreement with MH with 4, 8 and 12 levels, this agreement dropped to below chance levels as sample sizes decreased, suggesting that Logit requires further analysis before it can be recommended.

Summary and Conclusions

The present paper sought to examine two factors; 1) Whether changing the number of levels of matching on the matching criterion and, 2) whether changing the size and the ratio of reference to focal group members impacts on the performance of MH and Logit for detecting DIF. Clearly, as the sample sizes decreased, mean values for MH and Logit decreased and agreement between MH and Logit declined. In addition, as the number of levels of matching increased in the full sample condition, agreement between Logit and the various MH values also dropped. Monte-Carlo research results have suggested that "thin" matching yields the best results (Donoghue & Allen, 1991). Clearly in the present case (using empirical data) there is no unequivocal way of knowing which items exhibit "true" DIF. One control which could have been utilized would be to take random samples from the reference group, where one would be considered the reference group and the other the focal. This approach would allow for a baseline assessment of the number of items labeled biased as a function of Type I error.

The results of the present study, taken within the context of other current research, suggests several other areas needing

further study. There is a clear need for research to determine which DIF procedure is best for very small sample conditions. The present findings as well as other recent research (Mazor et. al, 1991) suggest that the MH procedure is somewhat sample size dependent. Therefore, other procedures should continue to be studied under this condition. There also remains a need to determine the extent to which the apparent sample size dependent nature of MH results is impacted by the ratio of focal to reference group members. Sensitivity to departures from a one to one ratio have not been systematically examined. The reliance on a statistical significance test may also serve to make interpretation difficult. There is a necessary concern about Type I error rates due to repeated significance testing (one per item), which to date has not been addressed. There is also concern about what significance levels are satisfactory to flag items. Given the above observed sensitivity to sample size, this problem is even more of a concern. Rank ordering on the basis of magnitude of the DIF value may be a more reliable index of item bias. Agreement between MH procedures generally increased as sample sizes and the magnitude of the indices decreased. There exists a need for a systematic examination of the factors responsible. Is this phenomenon due to the absolute size of the samples decreasing, to a decrease in the relative difference between groups (ratio of reference to focal group members approaches one to one), some interaction of the two, or another factor(s)? No study to date has addressed these issues. The impact of utilizing MH alpha rather

than χ^2 needs to be considered. While the expectation is that the two should provide parallel results, little research to date has examined the relationship between the two. Rather, researchers have typically relied upon one or the other. In addition, the number of levels of matching issue needs further study. With current practice at times resulting in testing organizations utilizing all possible scores as matching levels, there is a need to examine the impact of having empty or near-empty cells on obtained results. Finally, the use of an appropriate criterion for matching on ability requires further consideration. While researchers have decried the use of a criterion internal to the test in question (test score) for matching, there has been little research to date on the use of criteria external to the test for estimating ability. Schultz (1992) has begun to examine the use of non-test-based indices for matching on ability.

Table 1

Sample Mantel-Haenszel Format

	Correct(1)	Incorrect(0)	Total
Reference	A_j	B_j	nR_j
Focal	C_j	D_j	nF_j
Total	R_j	W_j	T_j

Table 2

Agreement (Kappa) Between Procedures

Math Items

	MH-8		MH-12		Logit		
Sample	full	2	full	2	full	1	2
MH-4	75*	52*	59*	52*	96**	23*	04
MH-8			96*	100**	75*		13
MH-12					56*		13

Verbal Items

	MH-8		MH-12		Logit		
Sample	full	2	full	2	full	1	2
MH-4	65**	81**	65**	81**	98**	45*	24*
MH-8			98**	100**	65*		06
MH-12					63*		06

Note: Presents Kappa for each contrast. N=60 items for math, 85 for verbal.

Reference group sizes are 7320 (full), 1000 (sample 1), and 500 (sample 2).

Focal group sizes are 791 (full) and 500 (samples 1 and 2).

**p<.01. *p<.05.

Table 3

Descriptive Statistics for Procedures Across Conditions

Math Items

Sample	Full			1			2		
	Mean	SD	N	Mean	SD	N	Mean	SD	N
MH-4	4.98	8.87	22	2.49	3.92	12	2.33	4.28	11
MH-8	4.10	7.28	21				1.93	3.62	6
MH-12	4.00	7.10	19				1.89	3.51	6
Logit	5.05	8.82	24	3.13	4.75	19	2.34	3.51	10

Verbal Items

Sample	Full			1			2		
	Mean	SD	N	Mean	SD	N	Mean	SD	N
MH-4	6.47	9.91	33	3.15	5.39	22	2.32	3.99	17
MH-8	5.33	8.40	31				2.14	3.40	16
MH-12	5.33	8.38	32				2.16	3.41	16
Logit	6.67	10.44	34	3.52	5.52	22	2.71	4.21	17

Note: Presented are means and standard deviations for χ^2 for MH and G^2 for Logit. Also presented are the number of items identified as exhibiting DIF at each condition.

Table 4

Correlations Between Procedures

Math Items

Sample	Full			Sample 1		Sample 2			
	MH-8	MH-12	Logit	MH-4	Logit	MH-4	MH-8	MH-12	Logit
Full									
MH-4	98**	97**	99**	83**	79**	75**	71**	70**	71**
MH-8		99**	98**	83**	78**	76**	74**	75**	71**
MH-12			97**	82**	77**	76**	74**	75**	71**
Logit				84**	80**	76**	71**	71**	71**
Sample 1									
MH-4					74**	94**	90**	89**	69*
Logit						74**	71*	70*	93**
Sample 2									
MH-4							98**	97**	72*
MH-8								99**	71*
MH-12									71*

(Table continues)

Table 4 (continued)

Verbal Items

Sample	Full			Sample 1		Sample 2			
	MH-8	MH-12	Logit	MH-4	Logit	MH-4	MH-8	MH-12	Logit
Full									
MH-4	96**	96**	99**	87**	82**	76**	69*	68*	73**
MH-8		99**	96**	88**	81**	76**	75**	75**	74**
MH-12			95**	87**	80**	76**	75**	76**	68*
Logit				88**	82**	78**	71**	69*	73**
Sample 1									
MH-4					68*	91**	88**	87**	61*
Logit						57*	52*	51*	94**
Sample 2									
MH-4							95**	94**	50*
MH-8								99**	48*
MH-12									47*

Note: Presents pearson product moment correlations for each contrast. N=60 items for math, 85 for verbal.

Reference group sizes are 7320 (full), 1000 (sample 1), and 500 (sample 2).

Focal group sizes are 791 (full) and 500 (samples 1 and 2).

**p<.01. *p<.05.

References

Bradley, R. T., Fitzpatrick, A. R., & Sykes, R. C. (1991). The effects of number of score groups and score group size on the Mantel-Haenszel Alpha. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, Il.

Camilli, G., & Smith, J. K. (1990). Comparison of the Mantel-Haenszel test with a Randomized and a Jackknife test for detecting biased items. Journal of Educational Statistics, 15, 53-67.

Cohen, J. (1960). A coefficient of agreement for nominal scales. Educational and Psychological Measurement. 20, 37-46.

Donoghue, J. R., & Allen, N. L. (1991). "Thin" versus "Thick" matching in the Mantel-Haenszel procedure for detecting DIF. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, Il.

Fienberg, S. E. (1977). The analysis of cross-classified categorical data. Cambridge, MA: The MIT Press.

Hambleton R. K., & Rogers, H. J. (1988). Detecting biased test items: Comparison of the IRT area and Mantel-Haenszel methods. Paper presented at the annual meeting of the American Educational Research Association, New Orleans LA.

Hambleton, R. K., Rogers, H. J., & Arrasmith, D. (1988). Identifying potentially biased items: A comparison of the Mantel-Haenszel statistic and several item response theory models. Laboratory of Psychometric and Evaluative Research Report No. 154. Amherst, MA: University of Massachusetts, School of Education.

Holland, P. W., Longford, N. T., & Thayer, D. T. (1991). Stability of the MH D-DIF statistics across populations. Manuscript submitted for publication.

Holland, P., & Thayer, D. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), Test validity. Hillside, NJ: Lawrence Erlbaum.

Kok, F. G., Mellenbergh, G. J., & Van Der Flier, H. (1985). Detecting experimentally induced item bias using the iterative logit method. Journal of Educational Measurement, 22, 295-303.

Mantel, N., & Haenszel, N. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. Journal of the National Cancer Institute, 22, 719-748.

Mazor, K. M., Clauser, B. E., and Hambleton, R. K. (1991). The effect of sample size on the functioning of the Mantel-Haenszel statistic. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, Il.

Mellenbergh, G. J. (1982). Contingency table models for assessing item bias. Journal of Educational Statistics, 7, 105-118.

Perlman, C. L., Bezruczko, N., Reynolds, R. J., Rice, W. K., & Schulz, E. M. (1988). The stability of four methods for estimating item bias. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Raju, N. S., Bode, S. K., & Larsen, V. S. (1989). An empirical assessment of the Mantel-Haenszel statistic for studying differential item performance. Applied Measurement in Education, 2, 1-13.

Schultz, M. T. A comparison of some recently proposed procedures for the detection of biased test items. Unpublished doctoral dissertation, Fordham University, 1992.

Schulz, E. M., Perlman, C., Rice, W. K., & Wright, B. D. (1989). An empirical comparison of Rasch and Mantel-Haenszel procedures for assessing item bias. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Shermis, M. D., & St. George, R. (1990). Item bias in mathematics achievement: The progressive achievement tests for mathematics. Paper presented at the annual meeting of the National Council for Educational Measurement, Boston, MA.

Siegel, S., & Castellan, N. J. (1988). Nonparametric statistics for the behavioral sciences (2nd ed.). New York: McGraw Hill.

SPSS-X User's Guide. (1988). Chicago: SPSS Inc.

Van Der Flier, H., Mellenbergh, G. J., Ader, H. J., & Wijn, M. (1984). An iterative item bias detection method. Journal of Educational Measurement, 21, 131-145.

Wise, S. L. (1987). Differential item difficulty indicators in small samples. Paper presented at the annual meeting of the American Educational Research Association, Washington DC.

Wright, D. J. (1986). An empirical comparison of the Mantel-Haenszel and Standardization methods for detecting differential item performance. (Statistical Report No. SR-86-99). Princeton, NJ: Educational Testing Service.